

This Page Is Inserted by IFW Operations  
and is not a part of the Official Record

## **BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning documents *will not* correct images,  
please do not report the images to the  
Image Problem Mailbox.**

## REMARKS

### I. Introduction

Receipt is acknowledged of a non-final office action dated February 26, 2004. In the action, the examiner rejected claims 30, 33 and 35 as allegedly non-enabled, claims 11 and 30-45 as allegedly failing to meet the written description requirement, and claims 11, 31, 32, 34 and 36-43 as allegedly anticipated by Ellis et al., *Nature Genetics*, 8:285-290 (1994).

Reconsideration of this application is respectfully requested.

### II. Status of the Claims

In this amendment, applicants amended claim 11 and added new claims 60-62. Support for the new claims can be found throughout the specification, and on page 12, lines 4-6, and page 13, lines 32-34. Upon entry of this amendment, claims 11, 30-45 and 60-62 will be under examination.

Because the foregoing amendments do not introduce new matter, entry thereof by the Examiner is respectfully requested.

### III. Claims Rejected under 35 U.S.C. § 112, 1<sup>st</sup> paragraph

#### A. Enablement Rejection

Claims 30, 33 and 35 were rejected under 35 U.S.C. § 112, first paragraph, as allegedly non-enabled. In particular, the examiner asserted that “the application does not identify a single disease condition where the expression of SEQ ID NO: 1 or biological activity of SEQ ID NO: 1 may be a relevant factor or useful for the diagnosis of that disorder.” Office action at 3. In addition, the examiner stated that “since the identification of SEQ ID NO: 1 as a cell surface marker protein relies solely on its ‘significant’ identity with Xg<sup>a</sup>, not on an actual elucidation of diseases or conditions in which SEQ ID NO: 1 expression or activity is relevant, diagnosis of any condition or disease based upon the expression of SEQ ID NO: 1 is not enabled.” *Id.* Applicants respectfully traverse this ground for rejection.

**1. The specification describes how to make and use the claimed antibodies in diagnosis of a condition or disease**

One of skill in the art would know which diseases or disorders are associated with the expression of a cell surface glycoprotein and, therefore, the claimed antibodies which bind the cell surface glycoprotein. Indeed, the specification describes that such disorders include “hematologic, karyotypic, and neuronal disorders.” Specification at 22. *See also*, specification at 31-32, which specifically identifies various disorders associated with the expression of cell surface glycoproteins (CSGP).

Furthermore, the specification provides that certain disorders are associated with increased CSGP expression, and “an antagonist of CSGP may be administered to a subject to treat or prevent a disorder associated with increased expression or activity of CSGP.” Specification at 23. A skilled artisan would know which disorders are associated with increased expression of CSGP or can readily assay CSGP expression levels, for example, based on the teachings of the present specification. *See* specification at 32-33. In addition, the claimed invention describes that in one embodiment, the antagonist is “an antibody which specifically binds CSGP” (specification at 23) and one of skill in the art at the time of filing would know how to make and use such an antibody based on the teachings in the present application. Specification at 24-26.

Likewise, the application provides guidance on identifying polypeptides and fragments that are suitable antigens for antibody diagnosis and therapy. *See* working examples I-IX on pages 35-41 of the specification. Thus, a skilled artisan would be able to identify a disease or condition where the expression of SEQ ID NO: 1 or fragment thereof is useful for the diagnosis of that disease.

**2. The specification sufficiently describes the function of the claimed polypeptides**

The examiner stated that “a function of Xg<sup>a</sup> has not been identified” and therefore, reliance on sequence similarity between SEQ ID NO: 1 and Xg<sup>a</sup> to elucidate function is not reasonable. *See* Office Action at 3. However, at the time of filing, a skilled artisan would know that Xg<sup>a</sup> is a blood group antigen which is an “immunologically defined marker protein[] found on the surface of red blood cells” and “provide[s] a means for classification of blood group

systems.” Specification at 1, lines 17-19. Furthermore, the clinical significance of the blood group systems are well known. *Id.* at 1, lines 26-27.

In addition, applicants respectfully draw the Examiner’s attention to Ellis et al., *Nature Genetics*, Vol. 8, November 1994 (“Ellis I”), Ellis et al., *Nature Genetics*, Vol. 6, April 1994 (“Ellis II”), and page 2 of “The Blood Group Antigen Facts Book,” copies of which were submitted to the US PTO with an IDS filed on November 30, 2001. The Ellis papers explain the importance of XG (formerly PBDX), stating that the XG region maps to the X-specific region adjacent to MIC2 and is therefore a marker for MIC2 (Ellis I, page 289, column 1, paragraph 2). MIC2 exhibits an erythrocyte-specific quantitative polymorphism characterized by either a high or low level of CD99 (Ellis II, page 285, column 1, paragraph 2). Therefore, XG (and the protein of the invention) is a marker for such polymorphisms.

Also, an unusual kind of cis-acting regulation exists (possibly at the transcriptional level) at or near this site which appears to be involved in, and linked to, various sex-linked diseases including Klinefelter’s syndrome (Ellis II, page 394, column 1, paragraph 1).

Further, “The Blood Group Antigen Facts Book,” at page 252 explicitly states that the XG blood group system “has helped define the mechanism responsible for various sex-chromosome aneuploidies.” Considering the degree of sequence similarity between SEQ ID NO:1 and XG, the polypeptide of SEQ ID NO:1 may be used as a marker for the XG blood group in the same way that XG is used.

Applicants also respectfully draw Examiner’s attention to the attached paper by Steven Brenner et al., *Proc. Natl. Acad. Sci. USA* (1998) 95:6073-6078 (Exhibit I). Through exhaustive analysis of a dataset of proteins with known structural and functional relationships, Brenner et al. have determined that 30% identity is a reliable threshold for establishing evolutionary homology between two sequences aligned over at least 150 residues, and that 40% identity is a reliable threshold when aligned over at least 70 residues (pages 6073 and 6076).

As the Examiner stated that “the PBDX protein taught by Ellis et al...is 92.3% identical to SEQ ID NO. 1 [of the present invention]” (Office Action at 3), this degree of similarity exceeds the threshold for reliable prediction of homology between two proteins. Thus, a

molecule with 90% sequence identity to the protein of SEQ ID NO:1 would then also be expected to have a similar function as a XG blood group system marker.

For at least these reasons, Applicants' claims are enabled and, therefore, withdrawal of this ground for rejection is respectfully requested.

**B. Written Description Rejection**

Claims 11 and 30-45 were rejected under 35 U.S.C. § 112, first paragraph, as allegedly failing to meet the written description requirement. While the examiner conceded that "an antibody to SEQ ID NO: 1 or immunogenic fragments thereof meet the written description requirement," the examiner asserted that "[s]ince [a]pplicant does not appear to have been in possession of the genus of polypeptides to which the instantly recited antibody specifically binds[,] [a]pplicant...does not appear to be in possession of the genus of antibodies specifically binding these polypeptides." Office Action at 5. The examiner also indicated that "[a]pplicant does not appear to have provided a description of which polypeptide sequences are 'naturally occurring.'" Office action at 4. Applicants respectfully traverse this ground for rejection.

**1. The specification adequately describes antibodies which bind naturally occurring sequences sharing at least 95% identity with SEQ ID NO: 1 and fragments thereof**

Antibodies which bind biologically active polypeptide fragments of SEQ ID NO: 1 and naturally occurring polypeptides sharing at least 90% sequence identity to SEQ ID NO. 1 are sufficiently described in the specification. A variant that is at least 90% identical to a sequence of known function is highly likely to share the same function such that one of skill in the art would believe that, more likely than not, the variant would function as the known protein.

Furthermore, methods for identifying these cell surface glycoproteins and assays for determining the function of a cell surface glycoprotein and for identifying molecules which interact with SEQ ID NO: 1 are described in the present specification. *See* specification at 35-40 and at 41-43, respectively. Therefore, a skilled artisan could readily identify biologically active peptide fragments of SEQ ID NO: 1 and antibodies that interact with such fragments. Nevertheless, in the interest of expediting prosecution, and without acquiescing to the examiner's

rejection, applicants amended claim 11 to recite “at least 95%” sequence identity. Support for this amendment can be found throughout the specification, and on page 14, line 5, in particular.

#### **IV. Rejections Under 35 U.S.C. § 102**

Claims 11, 31, 32, 34 and 36-43 were rejected under 35 U.S.C. § 102(b) as allegedly anticipated by Ellis et al., asserting that “Ellis teaches the making of polyclonal rabbit and monoclonal murine antibodies to PBDX by immunizing animals with the immunogenic N-terminal peptide...with a C-terminal cysteine residue attached to cross-link the peptide to a thyroglobulin carrier” and that “Ellis teaches that the antibodies bind to PBDX protein and identify it as the naturally-occurring Xg<sup>a</sup> blood group antigen.” Office Action at 5. Applicants respectfully traverse this ground for rejection.

##### **A. The presently claimed invention is not anticipated by Ellis et al.**

Ellis does not teach the claimed invention. Ellis describes making an antibody to a naturally occurring Xg<sup>a</sup> blood group antigen, but not an antibody that binds SEQ ID NO: 1 or a biologically active or immunologically active fragment of SEQ ID NO: 1. Further, the examiner stated that Xga shares 92.3% sequence identity with SEQ ID NO: 1. In the interest of expediting prosecution, applicants amended claim 11 to recite “at least 95%” sequence identity. Thus, Ellis does not teach making an antibody to a naturally occurring amino acid sequence at least 95% identical to the amino acid sequence of SEQ ID NO: 1.

In view of the foregoing arguments, it is respectfully requested that the present rejections be withdrawn.

### CONCLUSION

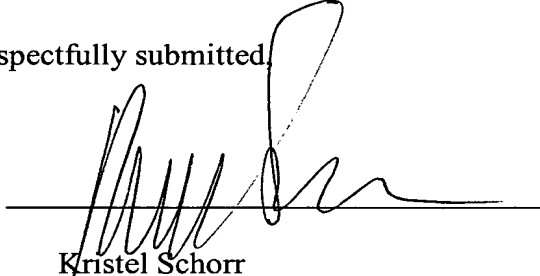
Reconsideration of the present application in view of the foregoing amendments and arguments is kindly requested.

It is respectfully urged that the present application is now in condition for allowance. Early notice to that effect is earnestly solicited.

Examiner Vandervegt is invited to contact the undersigned by telephone if it is felt that a telephone interview would advance the prosecution of the present application

Respectfully submitted,

By

  
Kristel Schorr  
Patent Agent  
Registration No. 55,600

Date: May 26, 2004

FOLEY & LARDNER LLP  
Washington Harbour  
3000 K Street, N.W., Suite 500  
Washington, D.C. 20007-5109  
Telephone: (202) 672-5574  
Facsimile: (202) 672-5399

Should additional fees be necessary in connection with the filing of this paper, or if a petition for extension of time is required for timely acceptance of same, the Commissioner is hereby authorized to charge Deposit Account No.19-0741 for any such fees; and applicant(s) hereby petition for any needed extension of time.

## Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER\*<sup>†‡</sup>, CYRUS CHOTHIA\*, AND TIM J. P. HUBBARD<sup>§</sup>

\*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and <sup>§</sup>Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)

**ABSTRACT** Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia C. (1995) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA  $ktup = 1$ , and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests have evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

**Previous Assessments of Sequence Comparison.** Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith–Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed ( $ktup = 2$ ) or greater effectiveness ( $ktup = 1$ ). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

© 1998 by The National Academy of Sciences 0027-8424/98/956073-6\$2.00/0  
PNAS is available online at <http://www.pnas.org>.

Abbreviation: EPQ, errors per query.

<sup>†</sup>Present address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126

<sup>‡</sup>To whom reprints requests should be addressed. e-mail: [brenner@hyper.stanford.edu](mailto:brenner@hyper.stanford.edu).



superfamilies. Pearson found that modern matrices and "ln-scaling" of raw scores improve results considerably. He also reported that the rigorous Smith–Waterman algorithm worked slightly better than FASTA, which was in turn more effective than BLAST.

Very large scale analyses of matrices have been performed (10), and Henikoff and Henikoff (11) also evaluated the effectiveness of BLAST and FASTA. Their test with BLAST considered the ability to detect homologs above a predetermined score but had no penalty for methods which also reported large numbers of spurious matches. The Henikoffs searched the SWISS-PROT database (12) and used PROSITE (13) to define homologous families. Their results showed that the BLOSUM62 matrix (14) performed markedly better than the extrapolated PAM-series matrices (15), which previously had been popular.

A crucial aspect of any assessment is the data that are used to test the ability of the program to find homologs. But in Pearson's and the Henikoffs' evaluations of sequence comparison, the correct results were effectively unknown. This is because the superfamilies in PIR and PROSITE are principally created by using the same sequence comparison methods which are being evaluated. Interdependency of data and methods creates a "chicken and egg" problem, and means for example, that new methods would be penalized for correctly identifying homologs missed by older programs. For instance, immunoglobulin variable and constant domains are clearly homologous, but PIR places them in different superfamilies. The problem is widespread: each superfamily in PIR 48.00 with a structural homolog is itself homologous to an average of 1.6 other PIR superfamilies (16).

To surmount these sorts of difficulties, Sander and Schneider (17) used protein structures to evaluate sequence comparison. Rather than comparing different sequence comparison algorithms, their work focused on determining a length-dependent threshold of percentage identity, above which all proteins would be of similar structure. A result of this analysis was the HSSP equation; it states that proteins with 25% identity over 80 residues will have similar structures, whereas shorter alignments require higher identity. (Other studies also have used structures (18–20), but these focused on a small number of model proteins and were principally oriented toward evaluating alignment accuracy rather than homology detection.)

A general solution to the problem of scoring comes from statistical measures (i.e., E-values and P-values) based on the extreme value distribution (21). Extreme value scoring was implemented analytically in the BLAST program using the Karlin and Altschul statistics (22, 23) and empirical approaches have been recently added to FASTA and SSEARCH. In addition to being heralded as a reliable means of recognizing significantly similar proteins (24, 25), the mathematical tractability of statistical scores "is a crucial feature of the BLAST algorithm" (1). The validity of this scoring procedure has been tested analytically and empirically (see ref. 2 and references in ref. 24). However, all large empirical tests used random sequences that may lack the subtle structure found within biological sequences (26, 27) and obviously do not contain any real homologs. Thus, although many researchers have suggested that statistical scores be used to rank matches (24, 25, 28), there have been no large rigorous experiments on biological data to determine the degree to which such rankings are superior.

**A Database for Testing Homology Detection.** Since the discovery that the structures of hemoglobin and myoglobin are very similar though their sequences are not (29), it has been apparent that comparing structures is a more powerful (if less convenient) way to recognize distant evolutionary relationships than comparing sequences. If two proteins show a high degree of similarity in their structural details and function, it

is very probable that they have an evolutionary relationship though their sequence similarity may be low.

The recent growth of protein structure information combined with the comprehensive evolutionary classification in the SCOP database (4, 5) have allowed us to overcome previous limitations. With these data, we can evaluate the performance of sequence comparison methods on real protein sequences whose relationships are known confidently. The SCOP database uses structural information to recognize distant homologs, the large majority of which can be determined unambiguously. These superfamilies, such as the globins or the immunoglobulins, would be recognized as related by the vast majority of the biological community despite the lack of high sequence similarity.

From SCOP, we extracted the sequences of domains of proteins in the Protein Data Bank (PDB) (30) and created two databases. One (PDB90D-B) has domains, which were all <90% identical to any other, whereas (PDB40D-B) had those <40% identical. The databases were created by first sorting all protein domains in SCOP by their quality and making a list. The highest quality domain was selected for inclusion in the database and removed from the list. Also removed from the list (and discarded) were all other domains above the threshold level of identity to the selected domain. This process was repeated until the list was empty. The PDB40D-B database contains 1,323 domains, which have 9,044 ordered pairs of distant relationships, or  $\approx 0.5\%$  of the total 1,749,006 ordered pairs. In PDB90D-B, the 2,079 domains have 53,988 relationships, representing 1.2% of all pairs. Low complexity regions of sequence can achieve spurious high scores, so these were masked in both databases by processing with the SEG program (27) using recommended parameters: 12 1.8 2.0. The databases used in this paper are available from <http://sss.stanford.edu/sss/>, and databases derived from the current version of SCOP may be found at <http://scop.mrc-lmb.cam.ac.uk/scop/>.

Analyses from both databases were generally consistent, but PDB40D-B focuses on distantly related proteins and reduces the heavy overrepresentation in the PDB of a small number of families (31, 32), whereas PDB90D-B (with more sequences) improves evaluations of statistics. Except where noted otherwise, the distant homolog results here are from PDB40D-B. Although the precise numbers reported here are specific to the structural domain databases used, we expect the trends to be general.

**Assessment Data and Procedure.** Our assessment of sequence comparison may be divided into four different major categories of tests. First, using just a single sequence comparison algorithm at a time, we evaluated the effectiveness of different scoring schemes. Second, we assessed the reliability of scoring procedures, including an evaluation of the validity of statistical scoring. Third, we compared sequence comparison algorithms (using the optimal scoring scheme) to determine their relative performance. Fourth, we examined the distribution of homologs and considered the power of pairwise sequence comparison to recognize them. All of the analyses used the databases of structurally identified homologs and a new assessment criterion.

The analyses tested BLAST (1), version 1.4.9MP, and WU-BLAST2 (2), version 2.0a13MP. Also assessed was the FASTA package, version 3.0t76 (3), which provided FASTA and the SSEARCH implementation of Smith–Waterman (8). For SSEARCH and FASTA, we used BLOSUM45 with gap penalties  $-12/-1$  (7, 16). The default parameters and matrix (BLOSUM62) were used for BLAST and WU-BLAST2.

**The "Coverage Vs. Error" Plot.** To test a particular protocol (comprising a program and scoring scheme), each sequence from the database was used as a query to search the database. This yielded ordered pairs of query and target sequences with associated scores, which were sorted, on the basis of their scores, from best to worst. The ideal method would have

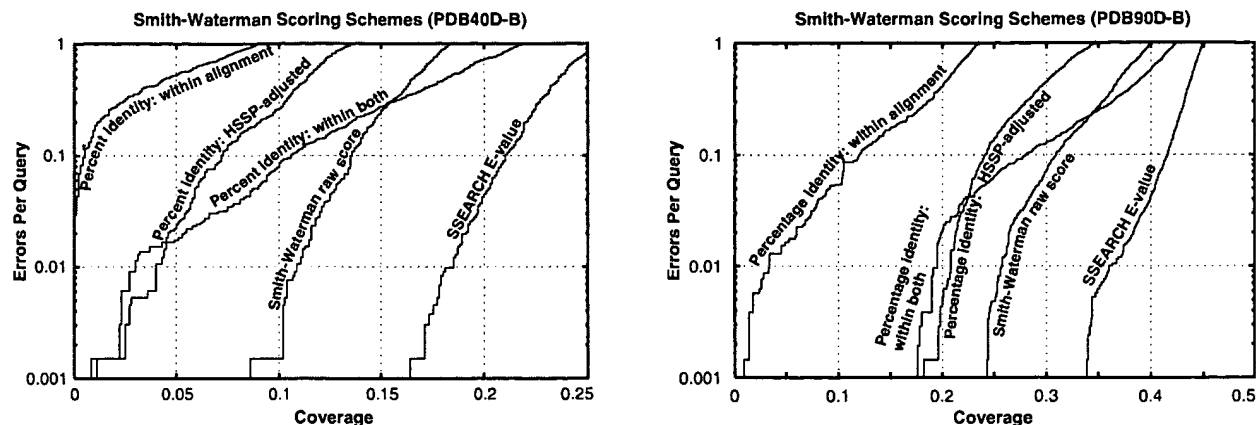


FIG. 1. Coverage vs. error plots of different scoring schemes for SSEARCH Smith-Waterman. (A) Analysis of PDB40D-B database. (B) Analysis of PDB90D-B database. All of the proteins in the database were compared with each other using the SSEARCH program. The results of this single set of comparisons were considered using five different scoring schemes and assessed. The graphs show the coverage and errors per query (EPQ) for statistical scores, raw scores, and three measures using percentage identity. In the coverage vs. error plot, the x axis indicates the fraction of all homologs in the database (known from structure) which have been detected. Precisely, it is the number of detected pairs of proteins with the same fold divided by the total number of pairs from a common superfamily. PDB40D-B contains a total of 9,044 homologs, so a score of 10% indicates identification of 904 relationships. The y axis reports the number of EPQ. Because there are 1,323 queries made in the PDB40D-B all-vs.-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The y axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The scores that correspond to the levels of EPQ and coverage are shown in Fig. 4 and Table 1. The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower right corner of the graph, which corresponds to identifying many evolutionary relationships without selecting unrelated proteins. Three measures of percentage identity are plotted. Percentage identity within alignment is the degree of identity within the aligned region of the proteins, without consideration of the alignment length. Percentage identity within both is the number of identical residues in the aligned region as a percentage of the average length of the query and target proteins. The HSSP equation (17) is  $H = 290.15l^{-0.562}$  where  $l$  is length for  $10 < l < 80$ ;  $H > 100$  for  $l < 10$ ;  $H = 24.7$  for  $l > 80$ . The percentage identity HSSP-adjusted score is the percent identity within the alignment minus  $H$ . Smith-Waterman raw scores and E-values were taken directly from the sequence comparison program.

perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice, perfect separation is impossible to achieve so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally determined homologs that have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of nonhomologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage vs. error plots, were devised to understand how

protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of Receiver Operating Characteristic (ROC) plots (33, 34) but better represent the high degrees of accuracy required in sequence comparison and the huge background of nonhomologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. Consistency is an aspect which has been largely

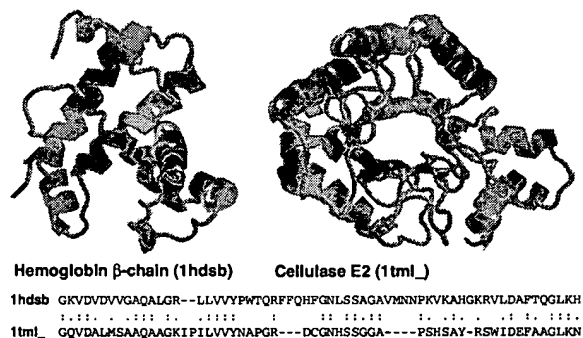


FIG. 2. Unrelated proteins with high percentage identity. Hemoglobin  $\beta$ -chain (PDB code 1hds chain b, ref. 38, *Left*) and cellulase E2 (PDB code 1tml, ref. 39, *Right*) have 39% identity over 64 residues, a level which is often believed to be indicative of homology. Despite this high degree of identity, their structures strongly suggest that these proteins are not related. Appropriately, neither the raw alignment score of 85 nor the E-value of 1.3 is significant. Proteins rendered by RASMOL (40).

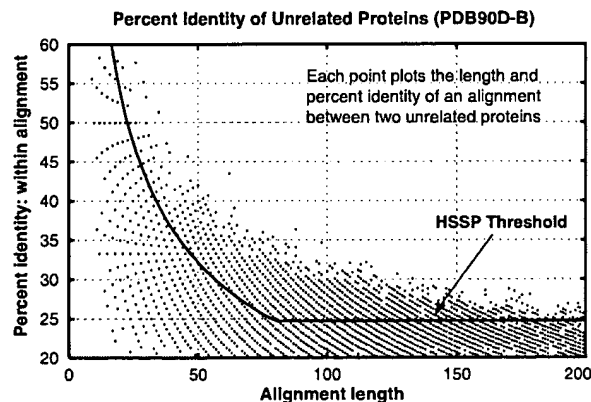


FIG. 3. Length and percentage identity of alignments of unrelated proteins in PDB90D-B: Each pair of nonhomologous proteins found with SSEARCH is plotted as a point whose position indicates the length and the percentage identity within the alignment. Because alignment length and percentage identity are quantized, many pairs of proteins may have exactly the same alignment length and percentage identity. The line shows the HSSP threshold (though it is intended to be applied with a different matrix and parameters).

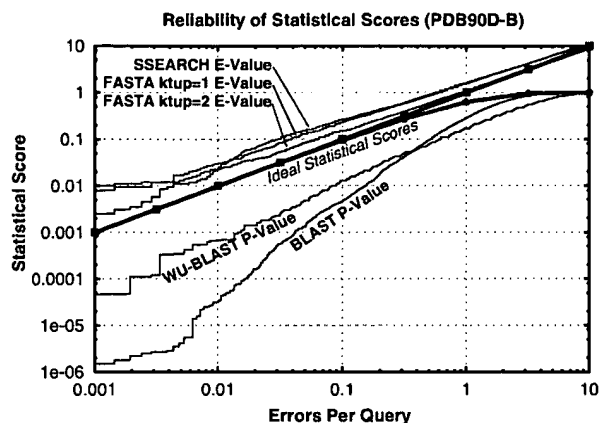


FIG. 4. Reliability of statistical scores in PDB90D-B: Each line shows the relationship between reported statistical score and actual error rate for a different program. E-values are reported for SSEARCH and FASTA, whereas P-values are shown for BLAST and WU-BLAST2. If the scoring were perfect, then the number of errors per query and the E-values would be the same, as indicated by the upper bold line. (P-values should be the same as EPQ for small numbers, and diverges at higher values, as indicated by the lower bold line.) E-values from SSEARCH and FASTA are shown to have good agreement with EPQ but underestimate the significance slightly. BLAST and WU-BLAST2 are overconfident, with the degree of exaggeration dependent upon the score. The results for PDB40D-B were similar to those for PDB90D-B despite the difference in number of homologs detected. This graph could be used to roughly calibrate the reliability of a given statistical score.

ignored in previous tests but is essential for the straightforward or automatic interpretation of sequence comparison results. Further, it provides a clear indication of the confidence that should be ascribed to each match. Indeed, the EPQ measure should approximate the expectation value reported by database searching programs, if the programs' estimates are accurate.

**The Performance of Scoring Schemes.** All of the programs tested could provide three fundamental types of scores. The first score is the percentage identity, which may be computed in several ways based on either the length of the alignment or the lengths of the sequences. The second is a "raw" or "Smith-Waterman" score, which is the measure optimized by the Smith-Waterman algorithm and is computed by summing the substitution matrix scores for each position in the alignment and subtracting gap penalties. In BLAST, a measure

related to this score is scaled into bits. Third is a statistical score based on the extreme value distribution. These results are summarized in Fig. 1.

**Sequence Identity.** Though it has been long established that percentage identity is a poor measure (35), there is a common rule-of-thumb stating that 30% identity signifies homology. Moreover, publications have indicated that 25% identity can be used as a threshold (17, 36). We find that these thresholds, originally derived years ago, are not supported by present results. As databases have grown, so have the possibilities for chance alignments with high identity; thus, the reported cutoffs lead to frequent errors. Fig. 2 shows one of the many pairs of proteins with very different structures that nonetheless have high levels of identity over considerable aligned regions. Despite the high identity, the raw and the statistical scores for such incorrect matches are typically not significant. The principal reasons percentage identity does so poorly seem to be that it ignores information about gaps and about the conservative or radical nature of residue substitutions.

From the PDB90D-B analysis in Fig. 3, we learn that 30% identity is a reliable threshold for this database only for sequence alignments of at least 150 residues. Because one unrelated pair of proteins has 43.5% identity over 62 residues, it is probably necessary for alignments to be at least 70 residues in length before 40% is a reasonable threshold, for a database of this particular size and composition.

At a given reliability, scores based on percentage identity detect just a fraction of the distant homologs found by statistical scoring. If one measures the percentage identity in the aligned regions without consideration of alignment length, then a negligible number of distant homologs are detected. Use of the HSP equation improves the value of percentage identity, but even this measure can find only 4% of all known homologs at 1% EPQ. In short, percentage identity discards most of the information measured in a sequence comparison.

**Raw Scores.** Smith-Waterman raw scores perform better than percentage identity (Fig. 1), but ln-scaling (7) provided no notable benefit in our analysis. It is necessary to be very precise when using either raw or bit scores because a 20% change in cutoff score could yield a tenfold difference in EPQ. However, it is difficult to choose appropriate thresholds because the reliability of a bit score depends on the lengths of the proteins matched and the size of the database. Raw score thresholds also are affected by matrix and gap parameters.

**Statistical Scores.** Statistical scores were introduced partly to overcome the problems that arise from raw scores. This scoring scheme provides the best discrimination between homologous proteins and those which are unrelated. Most

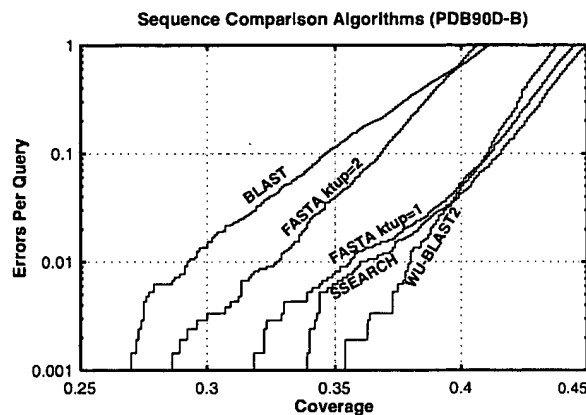
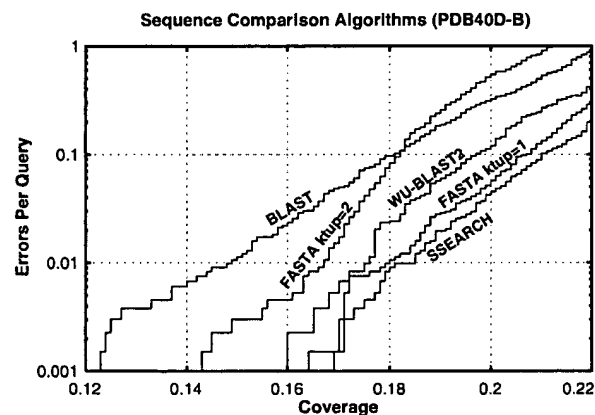


FIG. 5. Coverage vs. error plots of different sequence comparison methods: Five different sequence comparison methods are evaluated, each using statistical scores (E- or P-values). (A) PDB40D-B database. In this analysis, the best method is the slow SSEARCH, which finds 18% of relationships at 1% EPQ. FASTA ktup = 1 and WU-BLAST2 are almost as good. (B) PDB90D-B database. The quick WU-BLAST2 program provides the best coverage at 1% EPQ on this database, although at higher levels of error it becomes slightly worse than FASTA ktup = 1 and SSEARCH.

likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPQ for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPQ.

**Overall Detection of Homologs and Comparison of Algorithms.** The results in Fig. 5A and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPQ. BLAST, which identifies 15%, was the worst performer, whereas FASTA  $k_{\text{up}} = 1$  is nearly as effective as SSEARCH. FASTA  $k_{\text{up}} = 2$  and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA  $k_{\text{up}} = 1$ . WU-BLAST2 is slightly faster than FASTA  $k_{\text{up}} = 2$ , but the latter has more interpretable scores.

In PDB90D-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5B). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA  $k_{\text{up}} = 1$ , SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity

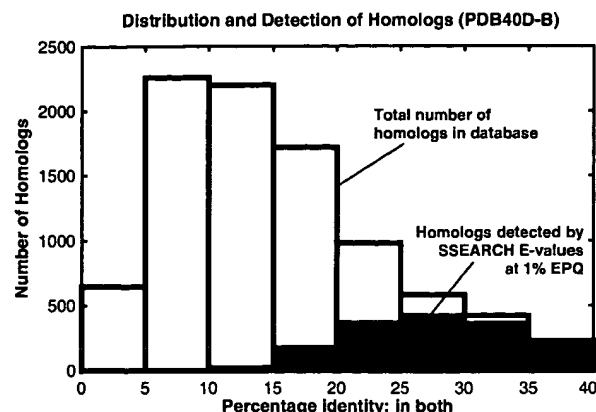


Fig. 6. Distribution and detection of homologs in PDB40D-B. Bars show the distribution of homologous pairs PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPQ. The PDB40D-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pairwise sequence comparison to detect them.

are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

## CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (i) using a large current database in which the protein sequences have been complexity masked and (ii) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB40D-B

Method	Relative Time*	1% EPQ Cutoff	Coverage at 1% EPQ
SSEARCH % identity: within alignment	25.5	>70%	<0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSSP-scaled	25.5	35% (HSSP + 9.8)	4.0
SSEARCH Smith-Waterman raw scores	25.5	142	10.5
SSEARCH E-values	25.5	0.03	18.4
FASTA $k_{\text{up}} = 1$ E-values	3.9	0.03	17.9
FASTA $k_{\text{up}} = 2$ E-values	1.4	0.03	16.7
WU-BLAST2 P-values	1.1	0.003	17.5
BLAST P-values	1.0	0.00016	14.8

\*Times are from large database searches with genome proteins.

extent of errors. Second, SSEARCH, WU-BLAST2, and FASTA ktup = 1 perform best, though BLAST and FASTA ktup = 2 detect most of the relationships found by the best procedures and are appropriate for rapid initial searches.

The homologous proteins that are found by sequence comparison can be distinguished with high reliability from the huge number of unrelated pairs. However, even the best database searching procedures tested fail to find the large majority of distant evolutionary relationships at an acceptable error rate. Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.\*\*

\*\*Additional and updated information about this work, including supplementary figures, may be found at <http://sss.stanford.edu/sss/>.

The authors are grateful to Drs. A. G. Murzin, M. Levitt, S. R. Eddy, and G. Mitchison for valuable discussion. S.E.B. was principally supported by a St. John's College (Cambridge, UK) Benefactors' Scholarship and by the American Friends of Cambridge University. S.E.B. dedicates his contribution to the memory of Rabbi Albert T. and Clara S. Bilgray.

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
- Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* **266**, 460–480.
- Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
- Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247**, 536–540.
- Brenner, S. E., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996) *Methods Enzymol.* **266**, 635–643.
- Pearson, W. R. (1991) *Genomics* **11**, 635–650.
- Pearson, W. R. (1995) *Protein Sci.* **4**, 1145–1160.
- Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147**, 195–197.
- George, D. G., Hunt, L. T. & Barker, W. C. (1996) *Methods Enzymol.* **266**, 41–59.
- Vogt, G., Etzold, T. & Argos, P. (1995) *J. Mol. Biol.* **249**, 816–831.
- Henikoff, S. & Henikoff, J. G. (1993) *Proteins* **17**, 49–61.
- Bairoch, A. & Apweiler, R. (1996) *Nucleic Acids Res.* **24**, 21–25.
- Bairoch, A., Bucher, P. & Hofmann, K. (1996) *Nucleic Acids Res.* **24**, 189–196.
- Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919.
- Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (National Bio-medical Research Foundation, Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345–352.
- Brenner, S. E. (1996) Ph.D. thesis. (University of Cambridge, UK).
- Sander, C. & Schneider, R. (1991) *Proteins* **9**, 56–68.
- Johnson, M. S. & Overington, J. P. (1993) *J. Mol. Biol.* **233**, 716–738.
- Barton, G. J. & Sternberg, M. J. E. (1987) *Protein Eng.* **1**, 89–94.
- Lesk, A. M., Levitt, M. & Chothia, C. (1986) *Protein Eng.* **1**, 77–78.
- Arratia, R., Gordon, L. & M, W. (1986) *Ann. Stat.* **14**, 971–993.
- Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268.
- Karlin, S. & Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 5873–5877.
- Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* **6**, 119–129.
- Pearson, W. R. (1996) *Methods Enzymol.* **266**, 227–258.
- Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* **12**, 215–226.
- Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* **266**, 554–571.
- Waterman, M. S. & Vingron, M. (1994) *Stat. Science* **9**, 367–381.
- Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13**, 669–678.
- Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Comm. Intl. Union Crystallogr., Cambridge, UK), pp. 107–132.
- Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997) *Curr. Opin. Struct. Biol.* **7**, 369–376.
- Orengo, C., Michie, A., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. (1997) *Structure (London)* **5**, 1093–1108.
- Zweig, M. H. & Campbell, G. (1993) *Clin. Chem.* **39**, 561–577.
- Gribskov, M. & Robinson, N. L. (1996) *Comput. Chem.* **20**, 25–33.
- Fitch, W. M. (1966) *J. Mol. Biol.* **16**, 9–16.
- Chung, S. Y. & Subbiah, S. (1996) *Structure (London)* **4**, 1123–1127.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25**, 3389–3402.
- Girling, R., Schmidt, W., Jr, Houston, T., Amma, E. & Huisman, T. (1979) *J. Mol. Biol.* **131**, 417–433.
- Spezio, M., Wilson, D. & Karplus, P. (1993) *Biochemistry* **32**, 9906–9916.
- Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* **20**, 374–376.